NTHU STAT 5410 Midterm Examination Nov 17, 2022

Instructions: Attempt all questions. Short and specific answers are preferred. Given explanation when required, but keep it as short and simple as possible. Give only one answer to each question – if you give alternative answers, the worst answer will be graded.

Question A If an egg were precisely an ellipsoid of revolution (i.e., ellipsoids with circular cross-section), we would expect its volume to be $V=LW^2/K$, where

- L is the long diameter,
- W is the diameter of the largest circular cross section,
- K is a constant $6/\pi$.

In a kitchen experiment, a researcher measured the diameters (L and W) of some hen's eggs with calipers and the volume V in the same way Archimedes had measured the volume of a golden crown, with a view to estimating the relation between these measures. From the directly observed variables, the data of three transformed variables $\mathrm{LogKV}=\log_{10}(\mathrm{KV})$, $\mathrm{LogL}=\log_{10}(L)$, and $\mathrm{LogW}=\log_{10}(W)$ were calculated and reported by the researcher. The results of some initial data analyses are given below, including the correlation matrix of them:

	LogL	LogW	LogKV
LogL	1.000	-0.257	0.398
LogW	-0.257	1.000	0.555
LogKV	0.398	0.555	1.000

and a brief descriptive summary of the three variables:

LogL		LogW		LogKV	
Min.	:0.735	Min.	:0.607	Min.	:1.96
1st Qu.	:0.752	1st Qu.	:0.616	1st Qu.	:2.00
Median	:0.766	Median	:0.622	Median	:2.00
Mean	:0.763	Mean	:0.621	Mean	:2.00
3rd Qu.	:0.773	3rd Qu.	:0.625	3rd Qu.	:2.01
Max.	:0.789	Max.	:0.636	Max.	:2.03

Output from a linear model, called `model A1`, with `LogKV` as the response and `LogL` and `LogW` as the predictors, i.e.,

$$E(\mathrm{LogKV}) = \beta_0 + \beta_1 \times \mathrm{LogL} + \beta_2 \times \mathrm{LogW}$$

is given below:

```
Coefficients:
             Estimate   Std. Error t value    Pr(>|t|)
(Intercept)  0.321      0.438      0.73       0.483
LogL         0.728      0.267      2.73       0.023
LogW         1.811      0.546      3.32       0.009
---
Multiple R-squared:  0.621
F-statistic: 7.38 on 2 and 9 DF,  p-value: 0.0127
```

```
Correlation of Coefficients:
     (Intercept)    LogL
LogL -0.66
LogW -0.89          0.26
```

(1) (1 pt) How many eggs on which the three variables were measured in this experiment?

(2) (1 pt) Explain why `LogL` and `LogW` has a *positive* correlation (i.e., `0.26`).

(3) (2 pts) Under `model A1`, perform a test of $H_0 : \beta_1 + \beta_2 = 3$ at significant level 0.05. You must report the value of the test statistic and explain how you get it, together with your conclusion (i.e., reject or not reject). **[Hint.** $t_9^{(0.975)} = 2.26,\ F_{1,9}^{(0.95)} = 5.12.$ **]**

(4) (1 pt) Suppose we are interested in the predictions of log volume on the following two settings of (`LogL`, `LogW`): (1, 0.85) and (0.75, 0.60), and would like to construct confidence intervals for predictions (either mean response or future observation) on the two settings. Which confidence interval would you expect to be wider? Explain.

Output from a model without intercept, i.e.,

$$E(\texttt{LogKV}) = \beta_1 \times \texttt{LogL} + \beta_2 \times \texttt{LogW},$$

called `model A2`, is given below:

```
Coefficients:
         Estimate   Std. Error t value    Pr(>|t|)
LogL     0.858      0.195      4.40       0.0013
LogW     2.168      0.240      9.05       3.9e-06
---
Multiple R-squared:    1
```

(5) (1 pt) Answer true or false to the statement: "because `model A2` has a much larger R^2 value than `model A1`, the former is a better fit than the latter", and explain why you chose true or false.

Let $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$ be the vectors of the residuals of the `models A1` and `A2`, respectively. Let $\hat{\epsilon}_3$ be the vector `LogKV`$-$`LogL`$-2\times$`LogW`.

(6) (2 pts) Can you rank the RSSs ($=|\hat{\epsilon}_i|^2$, i=1, 2, 3) obtained from the three sets of residuals from smallest to largest? If yes, do it and explain why you ranked them in this way. If no, explain why.

(7) (2 pts) With all the information available, can you find the sum of the second residuals $\hat{\epsilon}_2$? If yes, do it and explain how you get the answer. If no, explain why.

(8) (2 pts) We know that the degrees of freedom of $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$ are 9 and 10, respectively, so that $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$ respectively lie in a 9-dimensional and a 10-dimensional subspaces. When the third residuals $\hat{\epsilon}_3$ is treated as the residuals of a particular model, what is the degrees of freedom of $\hat{\epsilon}_3$ and what is this model? Explain.

(9) (1 pt) We usually use $\hat{\sigma}^2$=RSS/df to estimate σ^2, where df is the degrees of freedom of residuals. With all the analysis results available, can you rank the $\hat{\sigma}^2$'s calculated from $\hat{\epsilon}_1$, $\hat{\epsilon}_2$, $\hat{\epsilon}_3$ from smallest to largest? If yes, do it and explain why you ranked them in this way. If no, explain why.

Under `models A1` and A2, the 95% confidence regions of (β_1, β_2) are shown in Figure 1, panel (a) for `model A1` and panel (b) for `model A2`:

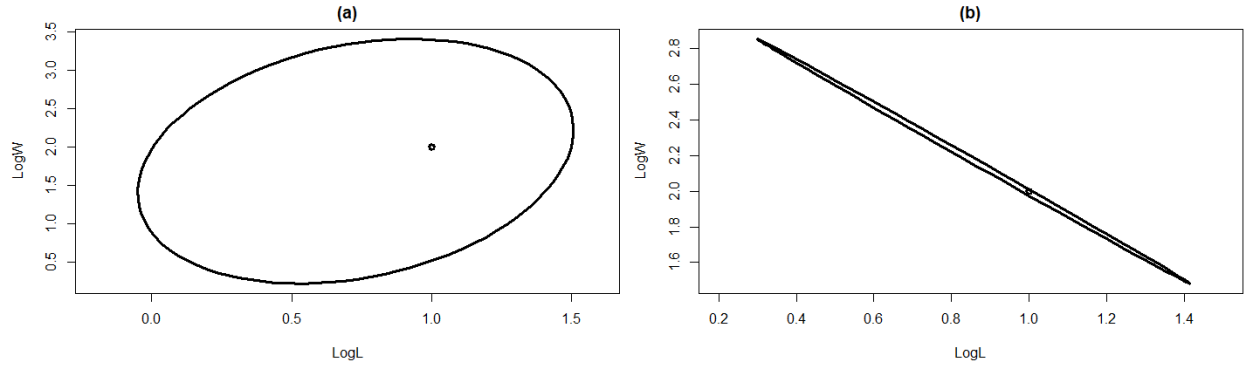


Figure 1

(10) (2 pts) In `model A2`, is the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ positive or negative? Explain how you got your answer, and why it is positive (or negative).

(11) (2 pts) Use all the analysis results available to test the assumption that hen's eggs are ellipsoids of revolution. You must present the model you choose and the null and alternative hypotheses, and explain how you get the test result (i.e., reject or not reject).

Question B Coleman report was an influential and controversial study, published by the US Government in 1966, under the title "Equality of Educational Opportunity." The report was based on an extensive survey of educational opportunity, was mandated in the Civil Rights Act of 1964, and was directed by the sociologist James Coleman. The national sample included almost 650,000 students and teachers in more than 3,000 schools However, the rate of nonresponse of schools to the survey was quite high. In fact, a disproportionately large number of big cities refused to participate in the survey. Coleman report was a landmark in policy research, being one of the first social scientific studies specifically commissioned by Congress in order to inform government policy. The research design adopted for the investigation changed the whole direction of policy research in education and was widely imitated by later researchers.

A random sample of information on 20 schools from the northeast and middle Atlantic states is drawn from the population of the Coleman Report (i.e., the 3,000 more schools). Its 5 variables are:

`Score:` verbal mean test score (all 6th graders),

`Salary:` staff salaries per pupil,

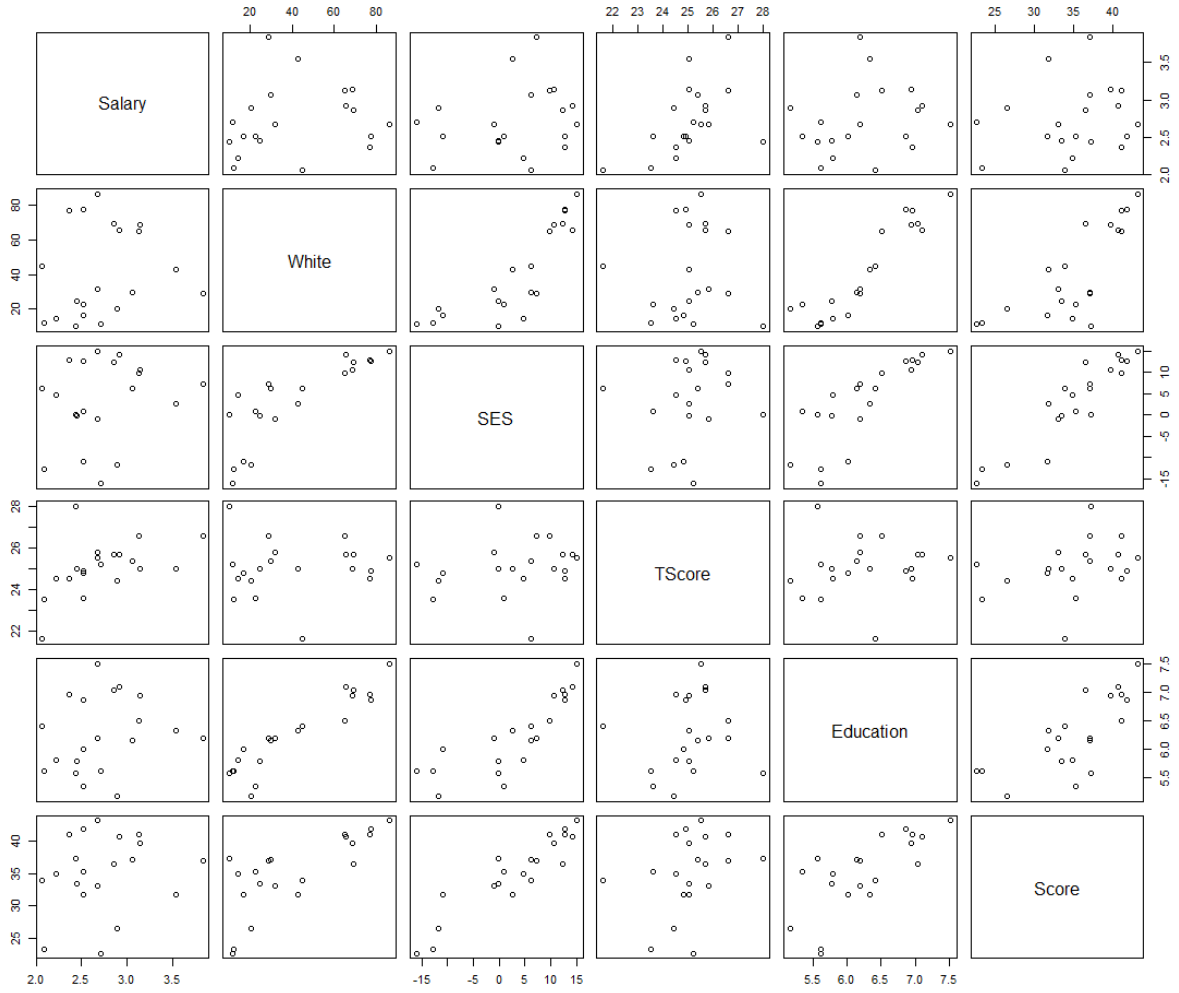
`White:` 6th grade per cent white-collar fathers,

`SES:` Socioeconomic status composite deviation: 6th grade means, for family size, family intactness, father's education, mother's education, per cent white collar fathers, and home items,

`TScore:` mean teacher's verbal test score,

`Education:` 6th grade mean mother's educational level (1 unit = 2 school years)

The results of some initial data analyses are given below, including scatter plots between these variables:



and the sample variances of the first 5 variables:

Salary	White	SES	TScore	Education
0.2062	670.73	92.65	1.726	0.4281

(12) (1 pt) In the scatter plots, which single variable predicts pupil's verbal performance best in this Coleman data? Explain.

A model, called `model B1`, with `Score` as the response and the other variables as the predictors is fit to this data its output is given below:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.9486	13.6275	1.46	0.165
Salary	-1.7933	1.2334	-1.45	0.168
White	0.0436	0.0533	0.82	0.427
SES	0.5558	0.0930	5.98	3.4e-05
TScore	1.1102	0.4338	2.56	0.023
Education	-1.8109	2.0274	-0.89	0.387

```
Residual standard error: 2.07 on 14 degrees of freedom
Multiple R-squared: ??
```

(13) (1 pt) Based on the *domain knowledge*, what is apparently surprising about some of the coefficient estimates in the fit of `model B1`? Explain.

(14) (2 pts) Explain why the scatter plot of `Score` and `Education` reveals a positive association between `Education` and `Score`, but the coefficient estimate of `Education` in the fit of `model B1` is negative.

(15) (1 pt) Interpret the coefficient estimate of `Education`, `-1.8109`, in the fit of `model B1`.

An alternative model, called `model B2`, with `Score` as the response and SES as the only predictor is fit to the data and its output is given below:

```
Coefficients:
             Estimate   Std. Error t value    Pr(>|t|)
(Intercept)  33.3228    0.5280     63.1       < 2e-16
SES          0.5603     0.0534     10.5       4.2e-09
---
Residual standard error: 2.24 on 18 degrees of freedom
Multiple R-squared:  0.86
```

(16) (1 pt) Based on the analysis results available, can you give the R-squared value of `model B1` (marked as ??) a reasonable non-zero lower bound *without doing any calculations*? If yes, explain why it is a lower bound. If no, explain why.

(17) (1 pt) What is the (sample) correlation of the data points in the scatter plot of SES and `Score`?

(18) (2 pts) Does the data support simplifying the model from `B1` to `B2`? Explain. **[Hint.** If $X \sim F_{a,b}$, $E(X) = b/(b-2)$. **]**

(19) (2 pts) This data is an observational data. Orthogonality is very unlikely to achieve in observational data. But, the coefficient estimates of SES in `models B1` and `B2` are almost identical. Explain why it happened.

(20) (2 pts) Suppose that the data supports `model B2`, in which SES explains a very high proportion (86%) of the variation in pupil's verbal performance. For the purpose of policy-making, can we claim that the other 4 predictors except SES have no impact on pupil's verbal performance? No matter your answer is yes or no, list at least 2 reasons to support your answer.